Abstract Title Page
Not included in page count.

Title: Coherent Power Analysis in Multi-Level Studies Using Design Parameters from Surveys.

Authors and Affiliations: Christopher Rhoads, University of Connecticut.

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Recent years have seen an increased interest in the use of randomized experiments to evaluate the causal effects of educational policies and practices in the United States (Angrist, 2004; Spybrook, Cullen and Lininger, 2011). Most large scale randomized experiments in education utilize the hierarchical structure of the U.S. educational system (students are nested in classrooms, classrooms are nested in schools, etc.) in the experimental design. Two basic types of designs are typically used. *Cluster randomized* or *hierarchical* trials (abbreviated CRT or HT) randomly assign entire clusters of students (such as schools) to a treatment group or a control group. *Randomized block* or *multi-site* trials (abbreviated RBD or MST) utilize clusters (such as schools) as blocks in the experimental design and randomly assign students within these blocks.

Researchers seeking funding to conduct a HT or MST will generally need to complete a power analysis in order to reassure the funder that the sample sizes (at the different levels of the design) will be sufficient to result in reasonable statistical power. Recent years have seen the development of software programs that facilitate the computation of statistical power when linear mixed models are used to analyze the types of cluster randomized and multi-site experiments described above. A recent article by Spybrook, Hedges and Borenstein (2014) describes two such programs: *Optimal Design Plus* (*OD Plus*, Raudenbush et. al., 2011) and *CRT Power* (Borenstein, Hedges and Rothstein, 2012).

In order to compute power in either of the two programs listed above the user must specify values for certain crucial design parameters. For instance, power for a two-level HT in *CRT Power* is computed after the user specifies the sample sizes at both levels of the hierarchy, the *effect size*, and the *intracluster correlation coefficient* (ICC). Power for a two-level MST in *CRT Power* is computed after the user specifies the sample sizes at both levels of the hierarchy, the *effect size*, the *intracluster correlation coefficient* (ICC) and an *effect size variance* parameter.

The need to specify an ICC and an effect size for the design of experiments in education has generated research seeking to clarify likely values for these parameters. These studies utilize survey based data sources (such as state longitudinal databases) to compute ICCs (eg. Hedges and Hedberg, 2007; Westine, Spybrook and Taylor, 2013) or benchmark effect sizes (eg. Bloom, Hill, Rebeck-Black and Lipsey, 2008; Scammacca, Fall and Roberts, 2015). Typically the ICCs and benchmark effect sizes are entered directly in to the power analysis software by researchers planning a MST or HT study.

There has been much less research on the topic of reasonable values for the effect size variance parameter. However, there is some evidence that researchers are beginning to pay more attention to obtaining empirical evidence about values of this parameter (Raudenbush, Reardon and Nomi, 2012; Raudenbush and Bloom, 2015).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The current paper notes that estimates of ICCs and/or effect sizes that come from surveys may need to be adjusted before being utilized to compute power for HTs or MSTs. Current software implementations prompt users to make assumptions about heterogeneity in treatment effects (effect size variance) only when performing a power analysis for MSTs. However, heterogeneity in treatment effects will impact the appropriate value of the ICC to use when planning a hierarchical trial. In particular, it is likely that ICCs computed from sample surveys will need to be adjusted to account for treatment effect heterogeneity when performing a power analysis for a hierarchical trial.

Similarly, the denominator of benchmark effect size values computed from sample surveys is likely to represent the variance of the outcome variable in the control group. When there are heterogeneous treatment effects, the variance in the treatment group is unlikely to equal the variance in the control group. Hence, benchmark effect size values may also need to be adjusted when used in a power analysis.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Previous work reporting ICCs and effect size benchmarks has failed to note the need for adjustments to these values in order to account for heterogeneity in treatment effects when planning a research study. Furthermore, writings about the computation of power in HT and MST designs typically define parameters solely in terms of "observed data" notation (eg. Raudenbush and Liu, 2000; Konstantopooulos, 2008). As a result the connection between these two sets of parameters has not been entirely clear. This has meant that a researcher comparing power for a multisite design with power for a hierarchical design could not be sure how assumptions about parameters made for one design should inform the assumptions made for the other design. For instance, should the ICC entered in to CRT Power when planning a two level HT be the same as the ICC which is entered when planning a two level MST?

The current paper resolves this issue using a potential outcomes approach. Writing parameters in terms of potential outcomes also makes clear how ICCs and/or effect sizes from surveys need to be adjusted to result in a correct power analysis for a HT or a MST. The result is a set of correction factors which researchers should use when planning future studies.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The paper assumes a design with *m* schools, each containing *2n* students. The statistical models considered are the usual linear mixed models used to define a power function for two level hierarchical and multisite trails. In particular, the model for the outcome score of the k^{th} student in the j^{th} school in the i^{th} treatment group in the hierarchical design is

$$Y_{ijk}^{H} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \ . \qquad (1)$$

The $\beta_{j(i)}$ parameters represent random effects associated with schools and the $\varepsilon_{ijk)}$ parameters represent student level random errors. The variance of the $\beta_{j(i)}$ parameters is $\sigma^2_{B,H}$, with the total

variance of the outcome measure in the hierarchical design given by $\sigma^2_{T,H}$. Let μ_E be the average value of the outcome variable when all students are assigned to the experimental condition and μ_C be the average value of the outcome variable when all students are assigned to the control condition. Then the effect size and ICC parameters necessary to correctly compute power in the hierarchical design are defined as follows: $\delta_h = \frac{\mu_E - \mu_c}{\sigma_{T,H}}$ and $\rho_H = \frac{\sigma^2_{B,H}}{\sigma^2_{T,H}}$. The model for the outcome score of the k^{th} student in the j^{th} school in the i^{th} treatment group in the multisite design is

$$Y^M_{ijk} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijk} \quad . \tag{2}$$

The γ_j parameters represent random variation in school means, the $\alpha\gamma_{ij}$ are random effects representing variation in treatment effects across schools with variance $\sigma^2_{\delta,M}/2$ and the ε_{ijk}) parameters represent student level random errors. Using the parameter definitions in equations (3)-(5) for the effect size, ICC and effect size variance will result in a correct power analysis for the multisite design. The symbol $\sigma^2_{B,C}$ refers to the between school variance in the control group and the symbol $\sigma^2_{W,C}$ refers to the within school variance in the control group.

$$\delta_M = \frac{\mu_E - \mu_c}{\sqrt{\sigma^2_{W,C} + \sigma^2_{B,C}}} \tag{3}$$

$$\rho_M = \frac{\sigma^2_{B,C}}{\sigma^2_{B,C} + \sigma^2_{W,C}} \tag{4}$$

$$\tau^2_2 = \frac{\sigma^2_{\delta,M}}{\sigma^2_{B,C} + \sigma^2_{W,C}} \tag{5}$$

We also write a standard two level hierarchical model in terms of potential outcomes as follows. The outcome that would be observed for the k^{th} student in the j^{th} school in the hypothetical world where everyone in the experiment was assigned to the control condition is modelled as $Y^C_{jk} = \mu_c + \beta^C_j + \varepsilon^C_{jk}, j = 1, \ldots, m; k = 1, \ldots, 2n.$ On the other hand, in the hypothetical world where everyone in the experiment was assigned to the experimental condition we would get the following model: $Y^E_{jk} = \mu_E + \beta^E_j + \varepsilon^E_{jk}, j = 1, \ldots, m; k = 1, \ldots, 2n.$ $\beta^C_j (\beta^E_j)$ are mean zero, random effects having variance $\sigma^2_{B,C} (\sigma^2_{B,E})$. The ε_{ijk} are mean zero random effects assumed to have the same variance, σ^2_W, in the experimental and control groups.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

The usefulness of the method is evident from the description in the Findings/Results section.

Findings / Results:

Description of the main findings with specific details.

Using the models defined above one can show the following relationship between the parameters in the hierarchical and multisite designs $\sigma^2_{B,H} = \sigma^2_{\delta,M}/2 + \sigma^2_{B,C} + \sigma_{\delta,\text{cov}}$. The $\sigma_{\delta,\text{cov}}$ symbol represents the covariance between treatment effects and school specific means in the control group. Because it is useful to write results in terms of a standardized version of this parameter we define $\tau_{\delta,\text{cov}} = \dfrac{\sigma_{\delta,\text{cov}}}{\sigma^2_{B,C} + \sigma^2_W}$.

ICCs computed from surveys describe an ICC defined in terms of control group quantities, namely, $\rho_s = \dfrac{\sigma^2_{B,C}}{\sigma^2_{B,C} + \sigma^2_W}$. Similarly, benchmark effect sizes from surveys represent a mean difference standardized by the total standard deviation in the control group, namely, $\delta_{bench} = \dfrac{\mu_T - \mu_c}{\sqrt{\sigma^2_{B,C} + \sigma^2_W}}$.

In the interest of space results are presented only for the case where one is planning a hierarchical trial using a survey based ICC and a benchmark effect size. In this case the ICC can be multiplied by a single correction factor that will simultaneously adjust for the impact of heterogeneity on the ICC and the impact of heterogeneity on the effect size to result in an accurate power analysis. This factor can be expressed as $k_b = 1 + \dfrac{n}{2n-1}\left(\dfrac{\tau^2_2 + 2\tau_{\delta,\text{cov}}}{\rho_S}\right)$. Clearly the correction factor depends on the covariance between school specific treatment effects and school specific control group means. The possible values for this parameter can be bounded using the covariance inequality. Table 1 in the appendix tabulates values of k_b for various values of τ^2_2, ρ_S and $\tau_{\delta,\text{cov}}$. The notation UB=Upper Bound and LB=Lower Bound is used.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Current practice for conducting power analyses in hierarchical trials using survey based ICC and effect size estimates may be misestimating power because ICCs are not being adjusted to account for treatment effect heterogeneity. Results presented in Table 1 show that the necessary adjustments can be quite large or quite small. Furthermore, power estimates without adjusting the ICC could be either too large or too small, depending on the covariance of school specific treatment effects with control group school means. The paper illustrates the need for the field to obtain better empirical evidence about likely values of this parameter in order to conduct accurate power analyses.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Angrist, J. (2004). American education research changes tack. *Oxford Review of Economic Policy, 20*(2), 198–212.

Bloom, H.S. and Weiland, C. (2015). Quantifying variation in Head Start effects on children's cognitive and socio-emotional skills using data from the National Head Start Impact Study. *MDRC Working Paper on Research Methodology*. New York: MDRC.

Bloom, H.,Hill, C.,Rebeck Black,A., and Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Efffectiveness, 1,* 289–328.

Borenstein, M., Hedges, L. V., & Rothstein, H. (2012). *CRT Power*. Teaneck, NJ: Biostat, Inc.

Hedges, L.V. & Hedberg, E.C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*, 66-88.

Raudnebush, S.W. and Bloom, H.S. (2015). Learning about and from variation in program impacts using multisite trials. *MDRC Working Paper on Research Methodology*. New York: MDRC.

Raudenbush, S.W. and Liu, X. (2000). Statistical analysis and optimal design for multisite randomized trials. *Psychological Methods,* 5, 199-213.

Raudnebush, S.W., Reardon, S. and Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness*, 5(3), 303-332.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). *Optimal Design Plus empirical evidence* (Version 3.0).

Scammacca, N., Fall, A.M. and Roberts, G. (2015). Benchmarks for expected annual academic growth for students in the bottom quartile of the normative distribution. *Journal of Research on Educational Effectiveness*, 8, 366-379.

Spybrook, J., Cullen, A., & Lininger, M. (2011). An examination of the impact of changes in federal policy on the landscape of education research. *Effective Education*, 3(2), 83-88.

Spybrook, J., Hedges, L., & Borenstein, M. (2014). Understanding statistical power in cluster randomized trials: Challenges posted by differences in notation and terminology. *Journal of Research on Educational Effectiveness*, 7, 384-406.

Westine, C, Spybrook, J. and Taylor, J.(2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review, 37(6), 490-519.*

Appendix B. Tables and Figures

Table 1: *Correction factor (k_b) when planning a hierarchical study with a survey based ICC and a benchmark effect size (n=25).*

		ρ_s					
τ_2^2	$\tau_{\delta,\text{cov}}$	0.001	0.01	0.05	0.1	0.2	0.3
0.01	LB	0.490	0.490	0.490	0.490	0.490	0.490
	LB/2	3.296	1.000	0.796	0.770	0.758	0.753
	0	6.102	1.510	1.102	1.051	1.026	1.017
	UB/2	8.908	2.020	1.408	1.332	1.293	1.281
	UB	11.714	2.531	1.714	1.612	1.561	1.544
0.05	LB	0.490	0.490	0.490	0.490	0.490	0.490
	LB/2	13.500	2.020	1.000	0.872	0.809	0.787
	0	26.510	3.551	1.510	1.255	1.128	1.085
	UB/2	39.520	5.082	2.020	1.638	1.446	1.383
	UB	52.531	6.612	2.531	2.020	1.765	1.680
0.1	LB	0.490	0.490	0.490	0.490	0.490	0.490
	LB/2	26.255	3.296	1.255	1.000	0.872	0.830
	0	52.020	6.102	2.020	1.510	1.255	1.170
	UB/2	77.786	8.908	2.786	2.020	1.638	1.510
	UB	103.551	11.714	3.551	2.531	2.020	1.850
0.2	LB	0.490	0.490	0.490	0.490	0.490	0.490
	LB/2	51.765	5.847	1.765	1.255	1.000	0.915
	0	103.041	11.204	3.041	2.020	1.510	1.340
	UB/2	154.316	16.561	4.316	2.786	2.020	1.765
	UB	205.592	21.918	5.592	3.551	2.531	2.190